Research Data Practice in the Life Sciences

Stuart Macdonald
EDINA National Data Centre
stuart.macdonald@ed.ac.uk

The full report 'Patterns of information use and exchange: case studies of researchers in the life sciences' is available at: http://www.rin.ac.uk/case-studies

Background:

Scientific advances, the availability of powerful new information and communications tools and services, and new policies governing research funding have brought major changes for life science researchers. Together these developments have significantly altered both their needs and their practices in acquiring, generating and using information resources.

Aims and methodology:

The broad aim of the RIN-funded Case Studies in Life Sciences project, undertaken by a team of social scientists and information specialists from the Institute for the Study of Science, Technology and Innovation (ISSTI) and from Information Services and the Digital Curation Centre (DCC) at the University of Edinburgh, was to improve understanding of information use and exchange in the life sciences and exchange.

Seven case studies were conducted across a diverse range of laboratories and research groups. They included groups studying humans, animals and plants, and covered different kinds of research context, encompassing analytical laboratory-based research, field research and in-silico research. The nature of the data used in the research process also varied, including quantitative, image, clinical, laboratory-derived and field data[1] (including aquaculture and botanic collections).

The starting point for each case study was the use of 'probes'- specially designed 'information lab books' – to chart individual researchers' information practices. We followed these up with detailed discussions in interviews and focus groups.

1 In some of the cases research is conducted primarily within the digital realm - the tools and instrumentation go beyond assisting the research (they are the research).

Key findings include:

- Sharing and exchanging information is central to the ethos of life science research, however individual researchers wish to choose what to share, with whom, and when
- Researchers used informal and trusted sources of advice from colleagues, rather than institutional service teams, to help identify information sources and resources
- The use of social networking tools for scientific research purposes was far more limited than expected
- Data and information sharing activities are mainly driven by needs and benefits perceived as most important by life scientists rather than 'top-down' policies and strategies
- Differences were apparent in the patterns of information use and exchange between different disciplines within the life scientists therein forcing the need to avoid generalisation in policy making

Research data sharing:

Technological developments in the Life Sciences (including genome-sequencing technology, microarray technology, improved mass-spectrometry techniques) have resulted in large scale generation of research data which needs to be stored, curated and analysed. The sharing of varied and complex datasets however is potentially more problematic than the sharing of research results via scholarly communications which remains the primary vehicle for dissemination and reward.

The evidence from the case studies highlighted the following:

- Researchers have concerns about misuse of research data, ethical restraints and IPR
- Some disciplines lend themselves more than others to 'openly' data sharing
- Data curation and/or sharing only becomes crucial at certain stages of research lifecycle
- Researchers retained a keen sense of 'ownership' and protectiveness towards data which represents their 'competitive advantage' and intellectual capital'
- Policies for data sharing may need to be driven by their demonstrated value rather than from adopting a one-size-fits-all framework
- Researchers felt only they had the subject knowledge to curate their own data

Data typology and taxonomy: differences and similarities between cases:

The research data used and created by the life science groups studied were categorised according to the range of research data detailed in the RIN publication 'Stewardship of Digital Research Data: a framework of principles and guidelines'[1] . The five categories are: *Scientific experiments*, data from lab equipment, often reproducible; *Models or simulations*, data generated from test models; *Observations*, specific phenomena recorded at a specific time or location, where the data constitutes a unique and irreplaceable record; *Derived data*, resulting from processing or combining "raw" or other data e.g. images, slides, graphs; *Canonical or reference data*, a collection of peer-reviewed data, most probably published and curated. Examples of the usage and creation of the above data types were noted across the case studies. It was found that data can generally be both input and output; can be re-purposed and re-positioned at more than one point on the research data lifecycle, dependent upon who uses them and how and why they are used.

In order to facilitate comparative analysis of the diversity of cases some form of taxonomic ordering was necessary. For example the international proteomic or genomic research programmes characterise high volume sharing of largely standardised (and thus homogenous) data. Systems biology, which attempts to pull together diverse data (e.g. genomic, gene expression, proteomic, metabolic data) is characterised by large-scale processing but of more heterogeneous information. It was not practicable to reduce the level of diversity of life science research in informational terms to two simple dimensions as empirical findings suggest that the heterogeneity of data comprises several elements that cannot readily be reduced to a single dimension. One source of heterogeneity is where judgement needs to be exercised over diverse data. Another type of heterogeneity is between different types of the same form of data such as graphical images (e.g. brain structures or cell differentiation). Issues of the level of interdisciplinarity, relative production cost and reproducibility are also at stake. We thus make reference to *heterogeneities*.

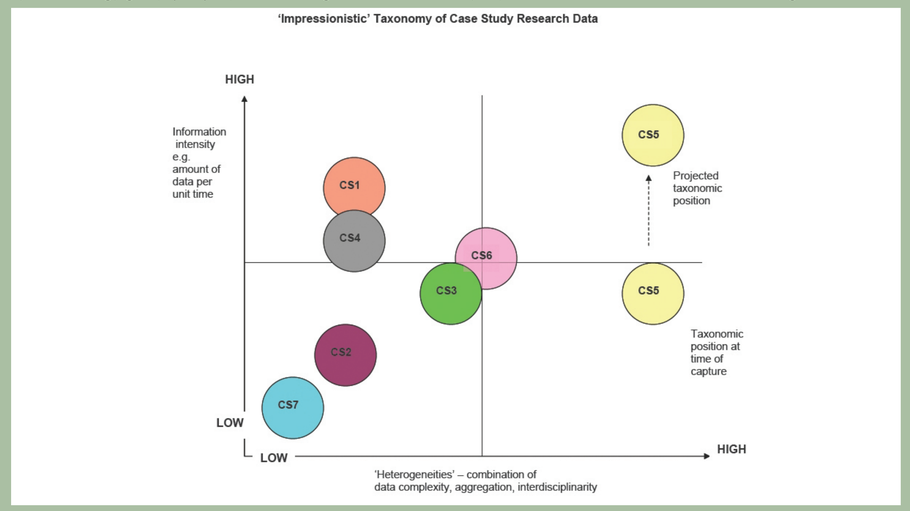
1 http://www.rin.ac.uk/data-principles

Case Study (CS)	Research Group	Research Data Type*
CS1	Animal Genetics and Animal Disease Genetics	**Scientific experiments**, Derived
CS2	Transgenesis in the Chick and Development of the Chick Embryo	**Scientific experiments**, Derived
CS3	Epidemiology of Zoonotic Diseases	**Observations**, Models/Simulations, Derived
CS4	Neuroscience	**Scientific experiments**, Observations, Derived
CS5	Systems Biology	**Models/Simulations**, Scientific experiments
CS6	Regenerative Medicine	**Scientific experiments**, Derived
CS7	Botantical Curation	**Observations**, Canonical

* the principal category is given first in each line

The 'Impressionistic' Taxonomy (below) provides an indication of approximate location of each case in terms of intensity and heterogeneities of research data exchange[2].

2 This diagram conveys relative rather than absolute positions. Locations are approximate. The arrow conveys how the Systems Biology group (CS5) expected to migrate from its currently experimental focus towards re-use of existing data.



The research team on this project comprised Robin Williams, Wendy Marsden, Ann Bruce, Jane Calvert from Innogen and ISSTI, Colin Neilson and Graham Pryor from DCC, and Stuart Macdonald from EDINA.